

Myopic Policy Bounds for Information Acquisition POMDPs

Mikko Lauri, Nikolay Atanasov, George J. Pappas, and Risto Ritala

Abstract—This paper addresses the problem of optimal control of robotic sensing systems aimed at autonomous information gathering in scenarios such as environmental monitoring, search and rescue, and surveillance and reconnaissance. The information gathering problem is formulated as a partially observable Markov decision process (POMDP) with a reward function that captures uncertainty reduction. Unlike the classical POMDP formulation, the resulting reward structure is nonlinear in the belief state and the traditional approaches do not apply directly. Instead of developing a new approximation algorithm, we show that if attention is restricted to a class of problems with certain structural properties, one can derive (often tight) upper and lower bounds on the optimal policy via an efficient myopic computation. These policy bounds can be applied in conjunction with an online branch-and-bound algorithm to accelerate the computation of the optimal policy. We obtain informative lower and upper policy bounds with low computational effort in a target tracking domain. The performance of branch-and-bounding is demonstrated and compared with exact value iteration.

I. INTRODUCTION

The current proliferation of sensors and robots has potential to transform fields as diverse as environmental monitoring, security and surveillance, localization and mapping, and structure inspection. One of the technical challenges in these scenarios is to control the sensors and robots to extract accurate information about various physical phenomena autonomously.

Robotic information acquisition may be thought of as an optimization problem: given constraints on robot motion and the available sensing resources, find an optimal control policy for the robot resulting in the greatest amount of collected information. Assuming Markovian system dynamics and observations conditionally independent given the system state, the optimization problem is formalized as a partially observable Markov decision process (POMDP) [1]. A POMDP consists of the state, action and observation spaces, stochastic system dynamics and observation models and a real-valued reward function. As the true state of the robot and environment are unknown, a probability density function (pdf) known as a belief state is maintained to represent information regarding the state. The optimal control policy is a mapping from belief states to actions such that the expected total reward accumulated over a given optimization horizon is maximized.

While POMDPs offer a principled framework handling probabilistic uncertainty in both control actions and sensing

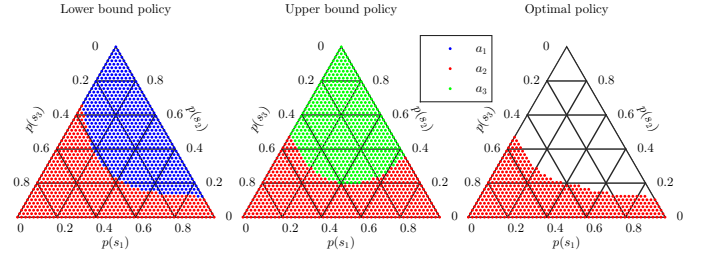


Fig. 1. The optimal policy μ^* in a POMDP (shown on the right) can be lower and upper bounded by myopic lower and upper bound policies μ_1^* and $\bar{\mu}_1^*$ (shown on the left and middle, respectively). The ternary plots cover the belief simplex \mathcal{B} in the three-state case with state space $\mathcal{S} = \{s_1, s_2, s_3\}$, while the dots indicate the action selected by the policy for a particular belief point. The inequalities $\mu_1^*(b) \leq \mu^*(b) \leq \bar{\mu}_1^*(b)$ hold for all $b \in \mathcal{B}$. The optimal policy is fully determined if $\mu_1^*(b) = \bar{\mu}_1^*(b)$. Even if the bounds do not agree they may provide information regarding the optimal action.

outcomes, obtaining optimal policies for them is computationally hard [2]. A wide range of algorithms for computing approximately optimal policies for POMDPs have been proposed in the literature, see e.g. [3], [4], [5] for reviews. Most research, however, focuses on a standard formulation in which the reward is a function of the hidden system state, and hence necessarily linear in the belief state. When it comes to information gathering applications in robotics (see e.g. [6], [7], [8], [9]), reward functions that capture uncertainty reduction (e.g. negative Shannon entropy [10]) or information gain (e.g., mutual information [10]) are often more interesting. These reward functions are nonlinear in the belief state, and hence, many of the existing approximation techniques for the standard POMDP formulation do not apply. Approximation methods applicable for such information acquisition POMDPs include some online algorithms [11] and specialized value iteration techniques, see e.g. [12]. Sampling-based online planning algorithms, such as those based on variants of Monte Carlo tree search (see e.g. [13], [14]), are also applicable.

Besides approximation algorithms, an alternative route to handling the inherent complexity of POMDPs with nonlinear belief-dependent rewards is to make structural statements about the optimal value function. In the case of state-dependent rewards, earlier work [15], [16], [17] shows that, if we restrict attention to a class of POMDPs that satisfy certain assumptions on the state and observation processes, we can bound the optimal policy from both below and above by an easily computable myopic (greedy) policy (see Fig. 1 for details). The specific case of POMDP multi-armed bandits was studied in [18].

The goal of this work is to extend these structural results to POMDPs with belief-dependent rewards. Krishnamurthy and Djonin [19] determined conditions under which the optimal

M. Lauri and R. Ritala are with Department of Automation Science and Engineering, Tampere University of Technology, P.O. Box 692, FI-33101, Tampere, Finland, {mikko.lauri, risto.ritala}@tut.fi.

N. Atanasov and G. Pappas are with GRASP Lab, University of Pennsylvania, Philadelphia, PA 19104, USA, {atanasov, pappasg}@seas.upenn.edu. This work was supported by TerraSwarm, one of six centers of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA.

policy has a threshold structure with respect to the monotone likelihood ratio (MLR) order, a partial order on the belief states. Such partial orders on the belief states have been used in the related active hypothesis testing field [20] to determine when a given sensing action dominates all other actions. For example, Naghshvar and Javidi [21] used Blackwell ordering of the observation models to reduce an active hypothesis test with K actions to a passive test with a single action. Compared to active hypothesis testing, information acquisition in POMDPs is even more challenging because the true underlying state evolves during the sensing process.

Our first contribution is to extend the structural results of [17] for POMDPs with state-dependent rewards to POMDPs with nonlinear belief-dependent rewards. This allows us to generate upper and lower bounds on the optimal policy (which agree in large portions of the belief state space and hence determine the optimal action) via a very efficient myopic policy computation. Our second contribution is to apply the myopic policy bounds in conjunction with an online branch-and-bound pruning algorithm to accelerate computation of the optimal policy. Our approach allows anytime computation with a bounded error from the optimal solution, unlike for instance sampling-based planning which is only asymptotically optimal.

The remainder of the paper is organized as follows. Section II formulates the information acquisition problem as a POMDP, and discusses suitable choices of reward functions. Section III briefly reviews existing structural results. Section IV introduces our extension to belief-dependent rewards and defines the myopic lower and upper bound policies. In Section V, we show how to apply the myopic policy bounds in a branch-and-bound pruning algorithm. Section VI presents empirical evaluation in target tracking domains, and Section VII concludes the paper.

II. INFORMATION ACQUISITION IN POMDPs

To model information acquisition problems in robotic applications we define a POMDP with a belief-dependent reward function. In detail, a POMDP is a tuple $\langle \mathcal{T}, \mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathbb{T}, \mathbb{O}, \rho \rangle$, where \mathcal{T} is a set of decision epochs and $\mathcal{S} = \{1, 2, \dots, S\}$, $\mathcal{A} = \{1, 2, \dots, A\}$, and $\mathcal{Z} = \{1, 2, \dots, Z\}$ are the finite state, action, and observation spaces, respectively. The function $\mathbb{T} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ is a stochastic state transition model such that $\mathbb{T}(s', s, a)$ is the probability of reaching state $s' \in \mathcal{S}$ from state $s \in \mathcal{S}$ after executing action $a \in \mathcal{A}$ and $\mathbb{O} : \mathcal{Z} \times \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ is a stochastic observation model such that $\mathbb{O}(z', s', a)$ is the conditional probability of observing $z' \in \mathcal{Z}$ when state $s' \in \mathcal{S}$ was reached after executing action $a \in \mathcal{A}$. To simplify notation, let $\mathbb{T}^a := [\mathbb{T}(j, i, a)]_{i,j=1}^S$ be the matrix of state transition probabilities for a given action $a \in \mathcal{A}$, i.e. previous states are column-wise and next states row-wise. Similarly, let $\mathbb{O}^a := [\mathbb{O}(z, j, a)]$ for $z \in \mathcal{Z}$, $j \in \mathcal{S}$, and $a \in \mathcal{A}$.

As information about the system state is incomplete, it is modeled by a probability density function (pdf) $b(s)$, $s \in \mathcal{S}$ over the system state. The set \mathcal{B} , called the belief space, is

the set of all possible pdfs over the system state. Finally, $\rho : \mathcal{B} \times \mathcal{A} \rightarrow \mathbb{R}$ is the expected reward, which is a (nonlinear) function of the belief state b .

Our definition subsumes the traditional POMDP definition with a state-and-action-dependent reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. This is seen defining $\rho(b, a) := \sum_{s \in \mathcal{S}} r(s, a)b(s)$.

The evolution of the belief state b in a POMDP is conditional on the actions and observations. When action $a \in \mathcal{A}$ is executed, the belief state evolves according to the state transition model \mathbb{T} . The evolution can be tracked by Bayesian filtering, which consists of iterating prediction and update steps. The prediction step revises the current belief state $b \in \mathcal{B}$ to the predicted belief $b^a \in \mathcal{B}$ according to

$$b^a(s') = \sum_{s \in \mathcal{S}} \mathbb{T}(s', s, a)b(s). \quad (1)$$

After the state transition, an observation $z' \in \mathcal{Z}$ is perceived and the information it provides is incorporated via the update step:

$$b'(s') = \tau(b, a, z') := \frac{1}{\eta(z' | b, a)} \cdot \mathbb{O}(z', s', a)b^a(s'), \quad (2)$$

where $\eta(z' | b, a) := \sum_{s \in \mathcal{S}} \mathbb{O}(z', s, a)b^a(s)$ is a normalization factor equal to the conditional probability of observing z' .

A. Reward functions for information acquisition

The goal in information acquisition is to reduce the uncertainty in the unobservable state of interest $s \in \mathcal{S}$. It is therefore desirable to reach “peaked” belief states that have the majority of their probability mass on a single underlying state. In other words, actions that lead the belief state towards the vertices of the probability simplex \mathcal{B} should be rewarded. Suitable measures of uncertainty are concave functions of the belief state, called uncertainty functions [22, Sec. 14.16].

Definition 1 (Uncertainty function and information gain [22]). *An uncertainty function is a non-negative, concave function $f : \mathcal{B} \rightarrow \mathbb{R}^+$. The information gain $\mathbf{I}_f(b, a)$ of an action $a \in \mathcal{A}$ in belief state $b \in \mathcal{B}$ is the expected reduction in the uncertainty function f :*

$$\mathbf{I}_f(b, a) := f(b^a) - \mathbb{E}_Z[f(\tau(b, a, Z))]. \quad (3)$$

Examples of uncertainty functions include Shannon entropy, the more-general Rényi quadratic entropy, and variants of the value of information, such as the probability of error in hypothesis testing [10]. For example, the information gain associated with Shannon entropy is known as mutual information.

For information acquisition tasks, either information gain or a negative uncertainty function may be used as the reward function in the POMDP formulation. However, as the following example shows, using an uncertainty function might be more appropriate when the magnitude of the predicted uncertainty $f(b^a)$ is significantly affected by the action choice.

Example 1. *Consider an active localization problem [23] in which a mobile robot needs to choose an appropriate action a to reduce the uncertainty in the distribution b of its current*

position s . Suppose that the entropy $f(b)$ of its current position distribution is n bits. The robot may choose between a risky high-velocity motion a_r leading to a predicted entropy of $f(b^{a_r}) = n + m$ bits or a safe low-velocity motion a_s leading to $f(b^{a_s}) = n + h$ bits, where $h < m$. Suppose that in both cases, after moving, the robot receives the same amount of information, k bits, from its sensor measurement z' . In this example, the negative uncertainty function correctly predicts that action a_r is risky:

$$\begin{aligned} -f(\tau(b, a_r, z')) &= -(n + m - k) \\ &< -(n + h - k) = -f(\tau(b, a_s, z)) \end{aligned}$$

but, perhaps surprisingly, the information gain prefers the risky action:

$$\mathbf{I}_f(b, a_r) = m - k > h - k = \mathbf{I}_f(b, a_s).$$

In the rest of the paper, we restrict attention to belief-dependent reward functions, specific to the task of information gathering, which have the following form:

$$\rho(b, a) := \sum_{s \in \mathcal{S}} r(s, a) b(s) - w_a f(b), \quad (4)$$

where f is an uncertainty function (preferred to information gain due to the observations in Example 1), $r(s, a)$ is any standard state-dependent reward function, and w_a are user-specified weights that trade-off measurement rewards and state uncertainty.

B. Value functions and optimal policies

Let $H \in \mathbb{N} \cup \{\infty\}$ denote a planning horizon so that $\mathcal{T} := \{1, 2, \dots, H\}$, and let $0 \leq \gamma \leq 1$ ¹ denote a discount factor determining the relative value of immediate and future rewards. The goal in information acquisition is to choose a policy $\mu_k^* : \mathcal{B} \rightarrow \mathcal{A}$ for each $k \in \mathcal{T}$ such that the expected sum of rewards over the decisions epochs is maximized. The sequence of optimal policies μ_k^* for $k \in \mathcal{T}$ remaining decisions can be computed via value iteration [24] according to:

$$Q_1^*(b, a) = \rho(b, a) \quad (5)$$

$$V_k^*(b) = \max_{a \in \mathcal{A}} Q_k^*(b, a), \quad (6)$$

$$Q_k^*(b, a) = \rho(b, a) + \gamma \sum_{z' \in \mathcal{Z}} \eta(z' | b, a) V_{k-1}^*(\tau(b, a, z')), \quad (7)$$

$$\mu_k^*(b) = \operatorname{argmax}_{a \in \mathcal{A}} Q_k^*(b, a), \quad (8)$$

where $V_k^* : \mathcal{B} \rightarrow \mathbb{R}$ is the optimal value function. For infinite-horizon problems with $H \rightarrow \infty$ and bounded rewards, value iteration converges to a unique fixed point V^* that satisfies the Bellman equation:

$$V^*(b) = \rho(b, a) + \gamma \sum_{z' \in \mathcal{Z}} \eta(z' | b, a) V^*(\tau(b, a, z')), \quad (9)$$

and the corresponding optimal infinite-horizon policy μ^* is stationary [24]. As the set \mathcal{B} of belief states over which

the value iteration must be solved is uncountable, the basic scheme above rarely results in practical methods for computing optimal policies.

III. MYOPIC POLICY BOUNDS FOR POMDPs

In this section, we review existing results on myopic policy bounds for POMDPs with state-dependent rewards. The idea of a myopic policy bound is that the optimal policy μ_k^* for any k can be either lower or upper bounded by the myopic policy μ_1^* (Eq. 8). The results exploit concepts of stochastic partial orders, summarized in the following definition.

Definition 2 (Stochastic partial orders [25]). Let $b_1, b_2 \in \mathcal{B}$.

- 1) b_1 first-order stochastically dominates b_2 , denoted $b_1 \geq_s b_2$, if $\sum_{i=j}^S b_1(i) \geq \sum_{i=j}^S b_2(i) \forall j \in \mathcal{S}$.
- 2) Equivalently, $b_1 \geq_s b_2$ iff $\sum_{i=1}^S b_1(i)g(i) \geq \sum_{i=1}^S b_2(i)g(i)$ for all increasing functions $g : \mathcal{S} \rightarrow \mathbb{R}$.
- 3) b_1 is greater than b_2 in the monotone likelihood ratio (MLR) order, denoted $b_1 \geq_r b_2$, if $b_1(i)b_2(j) \leq b_2(i)b_1(j)$, for all $i < j, j \in \mathcal{S}$.
- 4) A function $g : \mathcal{B} \rightarrow \mathbb{R}$ is MLR increasing (decreasing) if $b_1 \geq_r b_2 \Rightarrow g(b_1) \geq g(b_2)$ ($g(b_1) \leq g(b_2)$).
- 5) MLR dominance implies first-order stochastic dominance, $b_1 \geq_r b_2 \Rightarrow b_1 \geq_s b_2$.

Given conditions on the state transition and observation models, it was shown in [15], [16] that the optimal value function is MLR increasing and that the optimal policy may be lower bounded by μ_1^* . The required conditions, however, were thought to be too restrictive to be of practical significance until a recent extension [17] proposed the following revised assumptions.

Assumption 1 (Sufficient conditions for existence of myopic policy bounds [17]).

- (A1) \mathbb{T}^a and \mathbb{O}^a are totally positive or order 2 (TP2) $\forall a \in \mathcal{A}$,
- (A2) For all $j \in \{1, 2, \dots, S-1\}$, $a \in \{1, 2, \dots, A-1\}$, and $z \in \mathcal{Z}$, the matrices $D^{j,a,z} := [d_{m,n}^{j,a,z} + d_{n,m}^{j,a,z}] \in \mathbb{R}^{S \times S}$ are copositive, where $d_{m,n}^{j,a,z} := \mathbb{O}_{z,j}^a \mathbb{O}_{z,j+1}^{a+1} \mathbb{T}_{m,j}^a \mathbb{T}_{n,j+1}^{a+1} - \mathbb{O}_{z,j+1}^a \mathbb{O}_{z,j}^{a+1} \mathbb{T}_{m,j+1}^a \mathbb{T}_{n,j}^{a+1}$, and
- (A3) $\sum_{z \leq \bar{z}} \sum_{j \in \mathcal{S}} [\mathbb{T}_{i,j}^a \mathbb{O}_{z,j}^a - \mathbb{T}_{i,j}^{a+1} \mathbb{O}_{z,j}^{a+1}] \geq 0, \forall i \in \mathcal{S}, \forall \bar{z} \in \mathcal{Z}, \forall a \in \{1, 2, \dots, A-1\}$.

The TP2 property for a matrix requires that all its second-order minors are nonnegative [26]. Examples of TP2 observation models include exponential family distributions such as Gaussian, exponential, gamma, and binomial pdfs [26]; more examples may be found e.g. in [27]. While (A2) may be checked directly [28], it is much more efficient (but also more restrictive) to check the following:

- (A2') $d_{m,n}^{j,a,z} + d_{n,m}^{j,a,z} \geq 0$ for all $m, n \in \mathcal{S}, j \in \{1, \dots, S-1\}$, $a \in \{1, 2, \dots, A-1\}$, and $z \in \mathcal{Z}$.

Assumption (A3) is a revised version of the original assumption in [15], [16]. It states that for every state $i \in \mathcal{S}$, the prior pdf of perceiving $z \in \mathcal{Z}$ after ending in state $j \in \mathcal{S}$ after

¹Note that $\gamma = 1$ is only valid for finite H .

executing action $a + 1$ first order stochastically dominates that of action a . The assumptions lead to the following lemma.

Lemma 1 ([17]). Consider a POMDP $\langle \mathcal{T}, \mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathbb{T}, \mathbb{O}, \rho \rangle$. Then for all $b \in \mathcal{B}$, and $a', a \in \mathcal{A}$, $a' \geq a$,

- (L1) if (A3) holds, $\eta(z' | b, a') \geq_s \eta(z' | b, a)$, and
- (L2) if (A2) holds, $\tau(b, a', z') \geq_r \tau(b, a, z')$ for all $z' \in \mathcal{Z}$.

The following theorem for POMDPs with rewards linear in the belief state is due to [15, Prop. 2] and [17, Theorem 1]:

Theorem 1. Consider a POMDP $\langle \mathcal{T}, \mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathbb{T}, \mathbb{O}, \rho \rangle$ satisfying Assumptions (A1)-(A3), with $\rho(b, a) := \sum_{s \in \mathcal{S}} r(s, a) b(s)$. If $r(s, a)$ is increasing (decreasing) in S for all $a \in \mathcal{A}$, then the optimal value function V_k^* is MLR increasing (decreasing) and the optimal policy $\mu_k^*(b)$ for any $k \in \mathcal{T}$ satisfies $\mu_1^*(b) \leq \mu_k^*(b)$ ($\mu_1^*(b) \geq \mu_k^*(b)$) $\forall b \in \mathcal{B}$.

IV. TRANSFORMED MLR MONOTONE REWARDS FOR INFORMATION ACQUISITION POMDPs

In this section, we extend Theorem 1 to apply to reward functions of the form in (4), which are of interest for information acquisition tasks. To proceed, we need the following minor extension of Theorem 1.

Corollary 1. Consider a POMDP $\langle \mathcal{T}, \mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathbb{T}, \mathbb{O}, \rho \rangle$ satisfying Assumptions (A1)-(A3) and suppose that ρ is MLR increasing (decreasing) in \mathcal{B} for all $a \in \mathcal{A}$. Then, the optimal policy $\mu_k^*(b)$ for any $k \in \mathcal{T}$ satisfies $\mu_1^*(b) \leq \mu_k^*(b)$ ($\mu_1^*(b) \geq \mu_k^*(b)$) $\forall b \in \mathcal{B}$.

Proof. By [15, Prop. 2], it suffices to show V_k^* is MLR increasing. We proceed by induction. For the base case $k = 1$, $V_1^*(b) = \max_{a \in \mathcal{A}} \rho(b, a)$ is clearly MLR increasing. For the induction step, suppose the claim holds for V_{k-1}^* . Consider the sum part of the value iteration in Eq. (7). Let $b_1, b_2 \in \mathcal{B}$ s.t. $b_1 \geq_r b_2$. Now

$$\begin{aligned} & \sum_{z'=1}^Z \eta(z' | b_1, a) V_{k-1}^*(\tau(b_1, a, z')) \\ & \geq \sum_{z'=1}^Z \eta(z' | b_2, a) V_{k-1}^*(\tau(b_1, a, z')) \\ & \geq \sum_{z'=1}^Z \eta(z' | b_2, a) V_{k-1}^*(\tau(b_2, a, z')). \end{aligned} \quad (10)$$

The first inequality follows by Lemma (L1), [15, Lemma 1.2(2), 1.3(1)] which state $\tau(b, a, z')$ is MLR increasing in z' , and the induction hypothesis. The second inequality follows by Lemma (L2) and the induction hypothesis. The proof for the MLR decreasing part is similar and omitted. \square

Following the observation in Corollary 1, the main idea is to transform the reward function $\rho(b, a)$ (4) to one that is MLR monotone but leaves the optimal policy unchanged. Accordingly, we define two transformed reward functions:

$$\begin{aligned} \underline{\rho}(b, a) &= \rho(b, a) + [(I - \gamma(\mathbb{T}^a)^T)g]^T b \\ \bar{\rho}(b, a) &= \rho(b, a) + [(I - \gamma(\mathbb{T}^a)^T)h]^T b, \end{aligned} \quad (11)$$

where $g, h \in \mathbb{R}^S$ are parameter vectors chosen so that $\underline{\rho}(b, a)$ is MLR increasing and $\bar{\rho}(b, a)$ is MLR decreasing in \mathcal{B} .

Note that, by construction the transformed rewards in (11) leave the optimal policy unaffected. This can be verified by plugging either one into the Bellman equation (9). The corresponding infinite horizon value functions are $\underline{V}^*(b) = V^*(b) + g^T b$, $\bar{V}^*(b) = V^*(b) + h^T b$, and $\underline{\mu}^* \equiv \mu^* \equiv \bar{\mu}^*$ [17]. The following theorem gives necessary conditions for the existence of vectors g and h at a single belief state.

Theorem 2. Let $b_0 \in \mathcal{B}$. If there exists a $g \in \mathbb{R}^S$ such that the set of linear constraints given by

$$\left. \frac{\partial \rho(b, a)}{\partial b} \right|_{b=b_0} K + [(I - \gamma(\mathbb{T}^a)^T)g]^T K \geq 0 \quad \forall a \in \mathcal{A} \quad (12)$$

where K is an S -by- S matrix with entries

$$k_{ij} = \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } j = i + 1 \end{cases}, \quad (13)$$

are satisfied, then $\underline{\rho}(b, a)$ as defined in Eq. (11) is MLR increasing at b_0 .

Proof. As in the supplementary material of [17], let $\Delta = \{\delta \in \mathbb{R}^S : 1 = \delta(1) \geq \delta(2) \geq \dots \geq \delta(S)\}$. Any belief state $b \in \mathcal{B}$ may be represented as $\delta = Kb \in \Delta$. Let $b_1 = K\delta_1, b_2 = K\delta_2$ such that $b_1 \geq_r b_2$. By Definition 2, $b_1 \geq_s b_2$, and the equivalent partial order on $\Delta \subset \mathbb{R}^S$ is the component-wise partial order between δ_1 and δ_2 . Then, the function $\underline{\rho}(K\delta, a)$ is increasing on Δ if its partial derivatives are non-negative:

$$\begin{aligned} \frac{\partial \underline{\rho}(K\delta, a)}{\partial \delta} &= \frac{\partial \rho(K\delta, a)}{\partial \delta} + \frac{\partial [(I - \gamma(\mathbb{T}^a)^T)g]^T K \delta}{\partial \delta} \\ &= \frac{\partial \rho(b, a)}{\partial b} K + [(I - \gamma(\mathbb{T}^a)^T)g]^T K \geq 0 \end{aligned} \quad (14)$$

and the claim follows by evaluating the derivative at b_0 . \square

The proof for $\bar{\rho}(b, a)$ being MLR decreasing is obtained similarly, by swapping the required sign for the partial derivatives. Practically, the condition of the theorem may be checked by solving a feasibility linear program (LP) with $(S - 1) \cdot A$ constraints. If a single g and h may be chosen such that the transformed rewards are MLR monotone for all $b \in \mathcal{B}$, then the myopic policies

$$\begin{aligned} \underline{\mu}_1^*(b) &= \operatorname{argmax}_{a \in \mathcal{A}} \underline{\rho}(b, a) \\ \bar{\mu}_1^*(b) &= \operatorname{argmax}_{a \in \mathcal{A}} \bar{\rho}(b, a) \end{aligned} \quad (15)$$

satisfy $\underline{\mu}_1^*(b) \leq \mu^*(b) \leq \bar{\mu}_1^*(b)$ for all $b \in \mathcal{B}$. This follows from Corollary 1, and the fact that the optimal stationary policy is unaffected by applying the transformed rewards. It is important to note that the preceding inequalities do *not* hold anymore for finite horizon optimal policies.

Consider the following two examples of nonlinear belief-dependent reward functions obtained by applying different uncertainty functions in Eq. (4). In both examples, we denote by r_a the column vector formed from the elements of $r(s, a)$ for all $s \in \mathcal{S}$.

Example 2 (Shannon Entropy). *Let the uncertainty function in (4) be the Shannon entropy:*

$$f(b) := - \sum_{i=1}^S b(i) \log b(i).$$

The derivative of $\rho(b, a)$ for the constraints in Thm. 2 is:

$$\frac{\partial \rho(b, a)}{\partial b} = r_a^T + w_a [1 + \log(b(1)), \dots, 1 + \log(b(S))].$$

As $\log(0)$ is not well defined, we consider the inner belief simplex $\mathcal{B}_\epsilon := \{b \in \mathcal{B} \mid \forall i : b(i) \geq \epsilon\}$ for $\epsilon \in (0, 1)$. Note that

$$\log(\epsilon) \leq \log(b(i+1)) - \log(b(i)) \leq -\log(\epsilon).$$

For $i = 1, \dots, S-1$ and $\forall a \in \mathcal{A}$, the i th constraint in Eq. (12) can then be bounded by a belief-independent quantity:

$$\phi_i^a - \phi_{i+1}^a \geq w_a (\log(b(i+1)) - \log(b(i))) \geq -w_a \log(\epsilon),$$

where $\phi^a := r_a + (I - \gamma(\mathbb{T}^a)^T)g$.

Example 3 (Rényi Quadratic Entropy). *Let the uncertainty function in (4) be the Rényi quadratic entropy:*

$$f(b) := -\log \sum_{i=1}^S b^2(i).$$

The derivative of $\rho(b, a)$ for the constraints in Thm. 2 is:

$$\frac{\partial \rho(b, a)}{\partial b} = r_a^T + \frac{2w_a}{\sum_{i=1}^S b^2(i)} b^T.$$

Note that:

$$-1 \leq b(i+1) - b(i) \leq 1 \quad \text{and} \quad \frac{1}{S} \leq \sum_{i=1}^S b^2(i) \leq 1,$$

which can be used to obtain belief-independent bounds in (12). In particular, the i th constraint in (12) may be bounded by a belief-independent quantity:

$$\phi_i^a - \phi_{i+1}^a \geq \frac{2w_a}{\sum_{j=1}^S b^2(j)} (b(i+1) - b(i)) \geq -2w_a,$$

where $\phi^a := r_a + (I - \gamma(\mathbb{T}^a)^T)g$ and $i = 1, \dots, S-1$.

In both examples, belief independent bounds for finding h are obtained with straightforward changes. For Shannon entropy, one ideally wants to choose ϵ as small as possible to cover the greatest possible subset of the belief space. On the other hand, the smaller ϵ is, the more restrictive the constraints become, making it more unlikely to find feasible solutions. The issue is resolved by considering a different quantification of uncertainty, such as the Rényi quadratic entropy.

Figure 2 shows an example of the transformation of Rényi quadratic entropy rewards and the associated monotonicity properties as discussed in Example 3. The left axes show the original expected reward as function of the belief state, and the right axes show the transformed MLR increasing and decreasing rewards, respectively. For $S = 2$, the belief space \mathcal{B} can be represented by $[0, 1] \subset \mathbb{R}$ and MLR monotonicity

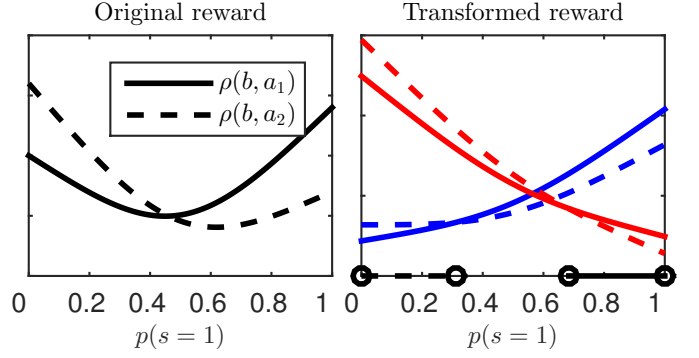


Fig. 2. Expected original (left axes) and transformed (right axes) rewards as function of the belief state according to Example 3. Rewards for action a_1 are shown by a solid line and for action a_2 by a dashed line. The increasing blue curves in the right axes depict the MLR increasing rewards $\rho(b, a)$, and the decreasing red curves depict the MLR decreasing rewards $\bar{\rho}(b, a)$. The two regions on the right axes indicated by lines starting and ending with circle markers are the parts of the belief space where the optimal policy is fully determined by the myopic policy bounds. The dashed region corresponds to $\mu_1^*(b) = \bar{\mu}_1^*(b) = a_2$ and the solid region to $\mu_1^*(b) = \bar{\mu}_1^*(b) = a_1$.

is equivalent to the familiar notion of monotonicity on \mathbb{R} . On the bottom of the right axes, two regions in \mathcal{B} are indicated by a dashed and a solid line. These are regions where the myopic policy bounds agree and the optimal policy is thus fully determined.

V. MYOPIC BOUNDS FOR BRANCH-AND-BOUND PRUNING

The significance of Theorems 1 and 2 is that they give a prescription for constructing policy bounds $\mu_1^*(b) \leq \mu^*(b) \leq \bar{\mu}_1^*(b)$ for all $b \in \mathcal{B}$. In this section, we show that the bounds can be used in an online branch-and-bound scheme to accelerate the computation of the optimal policy for POMDPs that satisfy Assumption 1.

Branch-and-bound pruning for POMDPs is based on a tree search over the set of belief states reachable by finite-length action and observation sequences, see e.g. [11]. The search is started from the current belief state b_0 at the root of the search tree. The tree branches due to action choices and observation possibilities, so that each node corresponds to a belief state b reachable from b_0 via the prediction and update iteration in (1), (2). The standard branch-and-bound algorithm works with lower and upper bounds on the optimal action values, i.e., $Q_k(b, a) \leq Q_k^*(b, a) \leq \bar{Q}_k(b, a)$. Starting from node b , the subtree corresponding to action a may be pruned if it is known to be suboptimal (i.e., the upper bound $\bar{Q}_k(b, a)$ is lower than the lower bound $Q_k(b, \hat{a})$ for some other action \hat{a}), thus saving computational resources.

Following the intuition of the branch-and-bound scheme above, we design a new algorithm (Alg. 1), which employs our policy bounds, *instead of* bounds on the optimal action values. Given a belief state b and a search depth $d > 0$, the algorithm constructs a search tree over reachable belief states, storing the estimate of the value $\hat{Q}(b, a)$ for each visited belief state. Only actions within the myopic lower and upper bound policies are considered.

Algorithm 1 Branch-and-bound with myopic policy bounds.**Input:** Belief state b , search depth d **Output:** Estimate of the value of the best action at b

```

1: function SEARCH( $b, d$ )
2:   if  $d = 1$  then return  $\max_a \rho(b, a)$ 
3:   end if
4:   for all  $a \in \mathcal{A} \cap [\mu_1^*(b), \bar{\mu}_1^*(b)]$  do
5:      $\hat{Q}(b, a) \leftarrow \rho(b, a) + \gamma \sum_{z' \in \mathcal{Z}} \left( \eta(z' | b, a) \right.$ 
6:        $\left. \times \text{SEARCH}(\tau(b, a, z'), d - 1) \right)$ 
7:   end for
8:   return  $\max_a \hat{Q}(b, a)$ 
9: end function

```

The next proposition follows from [11, Theorem 3.1].

Proposition 1. *After Alg. 1 terminates, its action recommendation \hat{a} is guaranteed to coincide with the optimal action $a^* := \mu^*(b)$ if the myopic lower and upper bounds agree at b , i.e., if $\mu_1^*(b) = \bar{\mu}_1^*(b)$. Otherwise, the suboptimality of \hat{a} is of order γ^d , i.e., $|Q^*(b, \hat{a}) - Q^*(b, a^*)| \sim o(\gamma^d)$.*

VI. APPLICATION TO TARGET TRACKING

We evaluate the performance of Algorithm 1 in a target tracking application. Let $s \in \mathcal{S}$ model the state of a physical process of interest such as the position of an adversarial aircraft or certain environment conditions (e.g., temperature, chemical concentration, air pollution, etc.). Let $s = 1$ correspond to an innocuous target state (e.g., the aircraft poses no danger or the environment levels are normal) and, at the other extreme, let $s = S$ correspond to a red-alert state in which there is imminent danger. We aim to design a control policy for an autonomous robotic platform to track the target state s . Let the available robot actions $\mathcal{A} = \{1, 2, \dots, A\}$ model the priority invested into tracking the target, with 1 corresponding to greatest priority (e.g., all available resources being dedicated to tracking the target) and A corresponding to lowest effort, i.e. the target is not being tracked.

The state transition model is such that the lower the priority invested into tracking, the more likely it is to transition to a state that is considered dangerous. Conversely, the higher the priority given to tracking, the less likely it is to accidentally enter a dangerous state. In other words, supposing the current state is i , the likelihood of transitioning to state $j > i$ should increase as function of a . State transition matrices \mathbb{T}^a modeling this can be defined e.g. as follows. Let $\{x_i\}_{i=1}^S$ and $\{y_j\}_{j=1}^{SA}$ be two increasing sequences of real numbers, and define a matrix $A_{ij} = \exp(x_i y_j)$, normalizing it such that each column sums to one, and define \mathbb{T}^a as the submatrix of A containing all rows and columns j for which $1 + (a-1) \cdot S \leq j \leq a \cdot S$. Each \mathbb{T}^a also satisfies the required TP2 condition [26].

The robot has access to $Z = S$ observations providing information about the target state. The observation matrix \mathbb{O}^a

for each action $a \in \mathcal{A}$ has elements

$$\mathbb{O}_{z,j}^a = \begin{cases} q & \text{if } z = j \\ (1-q)/2 & \text{if } j \neq 1, i \neq S \text{ and } |z-j| = 1 \\ (1-q) & \text{if } j = 1, i = S \text{ and } |z-j| = 1 \\ 0 & \text{otherwise} \end{cases}, \quad (16)$$

i.e. with probability $q < 1$ the observation identifies the true state, with symmetric error probability in either direction. This model corresponds to a situation where e.g. entering a dangerous state is perceived e.g. by the robot being damaged by hazardous environmental conditions. This class of target tracking domains can be checked to satisfy Assumptions 1.

A. Minimizing the uncertainty of the target state

We first consider target tracking with a penalty for uncertain tracker state information. We set $S = A = Z = 3$ and defined the sequences $\{x_i\} = \{1, 2, 3\}$, $\{y_j\} = \{-9, -8, \dots, -1\}$. The observation model was as defined in Eq. (16). The reward was as in Eq. (4), with

$$r(s, a) = \begin{bmatrix} 2 & 2.5 & 1 \\ 1.1 & 1.2 & 0.5 \\ 0.3 & 2 & 0.2 \end{bmatrix}, \quad (17)$$

$w_a = [1.1 \quad 1.6 \quad 1]$, and f equal to Rényi quadratic entropy. The discount factor was $\gamma = 0.99$, and the observation accuracy parameter was $q = 0.7$.

We computed the lower and upper myopic policy bounds for the problem. The bounds are visualized on the belief simplex \mathcal{B} in Figure 1. The figure shows on the left and middle the lower and upper myopic policies, and on the right the optimal policy determined when the two bounds agree.

We defined a larger problem with $S = Z = 8$ and $A = 3$, and in it compared Algorithm 1 against an exhaustive tree search. There are $N_d = (AZ)^d$ reachable belief states after d decisions, and the number of belief states in a complete search tree is $\sum_{i=0}^d N_i$.

TABLE I
AVERAGE PERCENTAGE OF BELIEF STATES PRUNED AS FUNCTION OF THE SEARCH DEPTH d APPLYING THE MYOPIC POLICY BOUNDS.

	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$
Belief states pruned	26.3%	36.3%	44.9%	52.3%	58.8%

Table I shows the average percentage of belief state pruned from the search tree for 100 randomly sampled initial belief states. Computing the myopic policy bound requires solving a feasibility LP with $(S-1) \cdot A$ constraints offline, and then during the search computing for each belief encountered the action maximizing the immediate expected transformed reward. Empirically, we found that for $d \geq 3$ branch-and-bounding was faster than the exhaustive search measured by computation time.

B. Comparison with existing approaches

Next, we compare the performance of our approach to an optimal incremental pruning algorithm² in a target tracking scenario with a state space size S between 4 and 256, and action space size A between 4 and 64. Since the incremental pruning algorithm (as well as other existing approaches) can handle only state-dependent rewards, we do not add an uncertainty function to the reward, i.e., $f \equiv 0$ in (4). The state-dependent reward $r(s, a)$ is designed to 1) incur a lower reward for investing higher priority in target tracking and to 2) penalize for both tracking the target poorly or too dangerously. The penalty on dangerous tracking models a situation where attempting to track the state too aggressively may make the robot vulnerable to environmental hazards. The reward is set as $r(s, a) = -c_a \cdot (A - a + 1) + t(s, a)$, where c_a is the cost of expending one unit of effort in tracking, and $t(s, a)$ is a tracking performance reward, defined

$$t(s, a) = \begin{cases} -c_p \cdot \frac{1}{s} & \text{if } s \in \mathcal{S}_p \\ -c_d \cdot \frac{1}{S-s+1} & \text{if } s \in \mathcal{S}_d \\ k_r \cdot s & \text{otherwise} \end{cases} \quad (18)$$

where \mathcal{S}_p and \mathcal{S}_d are the regions of the state space corresponding to poor and dangerous tracking, respectively, and c_p and c_d are the respective penalty costs. If the state is not in \mathcal{S}_l or \mathcal{S}_h , a tracking reward proportional to k_r is received.

In our experiments, we set $c_a = \frac{1}{2A}$, $c_p = 1$, $c_d = 0.1$, and $k_r = \frac{2}{S}$. Furthermore, we set $\mathcal{S}_l = \{s \in \mathcal{S} \mid s \leq S/5\}$ and $\mathcal{S}_h = \{s \in \mathcal{S} \mid s \geq 9S/10\}$, and $\gamma = 0.99$. The observation accuracy parameter was $q = 0.8$.

Table II shows the minimum, average and maximum percentage of actions that could be pruned by applying the myopic policy bounds as function of the domain size. A value of 0% would indicate that the bounds were completely non-informative, and a value of 100% that the optimal policy is fully determined by the bounds. The results were obtained by evaluating the bounds with 500 randomly sampled reachable belief states in each problem. We note that pruning efficiency tends to decrease both as function of the state space and action space size, but in all cases at least one third pruning rate was achieved even in the worst case. For $S = 256$ and $A = 64$, a vector satisfying the constraints of Thm. 2 could not be found.

As we computed the optimal solution, we could examine how close the bounds are to the optimal policy. Figure 3 shows an example of the looseness of the upper bound measured by the average of $\bar{\mu}_1^*(b) - \mu^*(b)$ over the 500 belief states. The greater the difference is, the further away the myopic upper bound policy is from the optimal policy, and the less informative it is. We note that even as the number of states and actions increases, the upper bound remains at a distance of less than 10 actions from the optimal.

²We applied the implementation from the pomdp-solve package of A. Cassandra, see <http://pomdp.org>

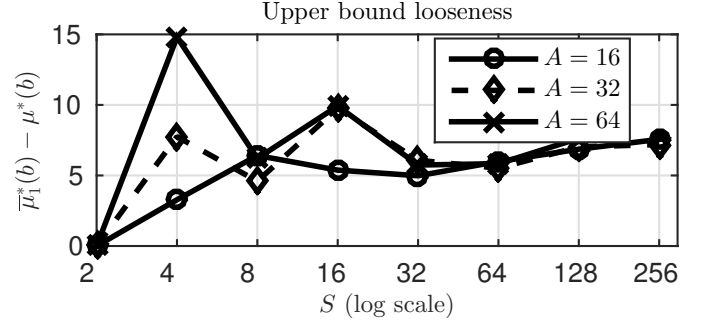


Fig. 3. The average distance between the upper bound and optimal policy as function of state space size S for various action space sizes A .

VII. CONCLUSION

We examined information acquisition in POMDPs with a reward function nonlinear in the belief state. We showed that if the POMDP fulfills certain structural properties, the optimal infinite horizon stationary policy may be lower and upper bounded by myopic (greedy) lower and upper bound policies. Based on the bounds, we designed a branch-and-bound pruning algorithm for online planning in POMDPs, and demonstrated its effectiveness in a target tracking application.

The main advantage of our approach is that, if the structural properties are satisfied in the POMDP model, the resulting branch-and-bound algorithm saves orders of magnitude in computation compared to the existing approaches. The main drawback is that the required structural properties can be quite restrictive in some applications. For example, a natural order in the state and action spaces is required.

Future work will focus on relaxing the structural requirements while keeping track of the effect on the policy bounds. This has potential to widen the scope of the possible applications of our approach significantly.

REFERENCES

- [1] L. Kaelbling, M. Littman, and A. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1-2, pp. 99-134, 1998. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S000437029800023X>
- [2] O. Madani, S. Hanks, and A. Condon, "On the Undecidability of Probabilistic Planning and Related Stochastic Optimization Problems," *Artificial Intelligence*, vol. 147, no. 1-2, pp. 5-34, 2003. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0004370202003788>
- [3] W. Lovejoy, "A Survey of Algorithmic Methods for Partially Observed Markov Decision Processes," *Annals of Operations Research*, vol. 28, no. 1, pp. 47-65, 1991.
- [4] M. Hauskrecht, "Value-function Approximations for Partially Observable Markov Decision Processes," *Journal of Artificial Intelligence Research*, vol. 13, no. 1, pp. 33-94, 2000.
- [5] G. Shani, J. Pineau, and R. Kaplow, "A Survey of Point-based POMDP Solvers," *Autonomous Agents and Multi-Agent Systems*, vol. 27, no. 1, pp. 1-51, 2013.
- [6] C. Stachniss, G. Grisetti, and W. Burgard, "Information Gain-based Exploration using Rao-Blackwellized Particle Filters," in *Proc. Robotics: Science and Systems (RSS)*, Cambridge, MA, USA, Jun. 2005.
- [7] B. Charrow, V. Kumar, and N. Michael, "Approximate Representations for Multi-robot Control Policies that Maximize Mutual Information," *Autonomous Robots*, vol. 37, no. 4, pp. 383-400, Aug. 2014.

TABLE II

THE (MINIMUM, AVERAGE, MAXIMUM) PERCENTAGE OF ACTIONS PRUNED IN TARGET TRACKING DOMAINS AS FUNCTION OF STATE SPACE SIZE S AND ACTION SPACE SIZE A .

S	$A = 4$	$A = 8$	$A = 16$	$A = 32$	$A = 64$
4	(75.0, 94.9, 100)	(75.0, 83.3, 87.5)	(75.0, 78.7, 87.5)	(71.9, 74.7, 75.0)	(64.1, 72.0, 75.0)
8	(50.0, 75.0, 75)	(37.5, 59.5, 62.5)	(50.0, 58.0, 75.0)	(59.4, 64.9, 71.9)	(48.4, 57.6, 67.2)
16	(50.0, 72.7, 75)	(50, 65.2, 75.0)	(43.8, 61.3, 68.8)	(46.9, 55.0, 62.5)	(46.9, 55.5, 62.5)
32	(50.0, 73.9, 75)	(37.5, 58.5, 62.5)	(37.8, 54.9, 62.5)	(37.5, 49.9, 59.4)	(43.8, 52.6, 59.4)
64	(50.0, 71.9, 75)	(37.5, 60.4, 62.5)	(43.8, 54.0, 62.5)	(37.5, 49.6, 59.4)	(40.6, 47.9, 53.1)
128	(50.0, 73.2, 75)	(50, 61.1, 62.5)	(43.8, 54.7, 56.3)	(40.6, 49.3, 53.1)	(42.2, 48.3, 53.1)
256	(50.0, 72.8, 75)	(50, 60.4, 62.5)	(43.8, 50.9, 56.3)	(43.8, 49.1, 53.1)	(n/a)

- [8] N. Atanasov, J. Le Ny, K. Daniilidis, and G. J. Pappas, "Information Acquisition with Sensing Robots: Algorithms and Error Bounds," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, Jun. 2014, pp. 6447–6454.
- [9] M. Lauri and R. Ritala, "Optimal Sensing via Multi-armed Bandit Relaxations in Mixed Observability Domains," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, WA, May 2015, pp. 4807–4812.
- [10] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.
- [11] S. Ross, J. Pineau, S. Paquet, and B. Chaib-draa, "Online planning algorithms for POMDPs," *Journal of Artificial Intelligence Research*, vol. 32, no. 1, pp. 663–704, 2008. [Online]. Available: <http://www.aaai.org/Papers/JAIR/Vol32/JAIR-3217.pdf>
- [12] M. Araya-López, O. Buffet, V. Thomas, and F. Charpillet, "A POMDP Extension with Belief-dependent Rewards," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., Vancouver, Canada, Dec. 2010, pp. 64–72.
- [13] D. Silver and J. Veness, "Monte-Carlo Planning in Large POMDPs," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., Vancouver, Canada, Dec. 2010, pp. 2164–2172.
- [14] A. Somani, N. Ye, D. Hsu, and W. S. Lee, "DESPOT: Online POMDP Planning with Regularization," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 1772–1780. [Online]. Available: <http://papers.nips.cc/paper/5189-despot-online-pomdp-planning-with-regularization.pdf>
- [15] W. S. Lovejoy, "Some Monotonicity Results for Partially Observed Markov Decision Processes," pp. 736–743, 1987.
- [16] U. Rieder, "Structural results for partially observed control models," *ZOR Zeitschrift für Operations Research Methods and Models of Operations Research*, vol. 35, no. 6, pp. 473–490, Nov. 1991. [Online]. Available: <http://link.springer.com/10.1007/BF01415990>
- [17] V. Krishnamurthy and U. Pareek, "Myopic Bounds for Optimal Policy of POMDPs: An Extension of Lovejoys Structural Results," *Operations Research*, vol. 63, no. 2, pp. 428–434, Apr. 2015. [Online]. Available: <http://pubsonline.informs.org/doi/10.1287/opre.2014.1332>
- [18] V. Krishnamurthy and B. Wahlberg, "Partially Observed Markov Decision Process Multiarmed Bandits—Structural Results," *Mathematics of Operations Research*, vol. 34, no. 2, pp. 287–302, 2009.
- [19] V. Krishnamurthy and D. V. Djonin, "Structured threshold policies for dynamic sensor scheduling - A partially observed Markov decision process approach," *IEEE Transactions on Signal Processing*, vol. 55, no. 10, pp. 4938–4957, Oct. 2007. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4305446>
- [20] M. Naghshvar and T. Javidi, "Active sequential hypothesis testing," *The Annals of Statistics*, vol. 41, no. 6, pp. 2703–2738, Dec. 2013. [Online]. Available: <http://projecteuclid.org/euclid.aos/1387313387>
- [21] —, "Active M-ary sequential hypothesis testing," in *2010 IEEE International Symposium on Information Theory*, vol. 1, no. 1. IEEE, Jun. 2010, pp. 1623–1627. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5513381>
- [22] M. H. DeGroot, *Optimal Statistical Decisions*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2004, Wiley Classics Library edition.
- [23] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. Cambridge, MA: The MIT Press, 2006.
- [24] D. Bertsekas and S. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 1996.
- [25] M. Shaked and J. G. Shanthikumar, *Stochastic Orders*. New York, NY: Springer New York, 2007. [Online]. Available: <http://link.springer.com/10.1007/978-0-387-34675-5>
- [26] S. M. Fallat and C. R. Johnson, *Totally Nonnegative Matrices*. Princeton, NJ: Princeton University Press, 2011.
- [27] S. Karlin, *Total Positivity*. Stanford University Press, 1968.
- [28] W. Kaplan, "A Test for Copositive Matrices," *Linear Algebra and its Applications*, vol. 313, no. 1–3, pp. 203–206, 2000.